Nucleic Acids Research, 2005, Vol. 33, Web Server issue W717–W723
doi:10.1093/nar/gki454

GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists

Marco Masseroli*, Osvaldo Galati and Francesco Pinciroli

BioMedical Informatics Laboratory, Bioengineering Department, Politecnico di Milano, I-20133 Milano, Italy

Received February 14, 2005; Revised and Accepted April 1, 2005

ABSTRACT

Phenotype analysis is commonly recognized to be of great importance for gaining insight into genetic interaction underlying inherited diseases. However, few computational contributions have been proposed for this purpose, mainly owing to lack of controlled clinical information easily accessible and structured for computational genome-wise analyses. We developed and made available through GFINDer web server an original approach for the analysis of genetic disorder related genes by exploiting the information on genetic diseases and their clinical phenotypes present in textual form within the Online Mendelian Inheritance in Man (OMIM) database. Because several synonyms for the same name and different names for overlapping concepts are often used in OMIM, we first normalized phenotype location descriptions reducing them to a list of unique controlled terms representing phenotype location categories. Then, we hierarchically structured them and the correspondent genetic diseases according to their topology and granularity of description, respectively. Thus, in GFINDer we could implement specific Genetic Disorders modules for the analysis of these structured data. Such modules allow to automatically annotate user-classified gene lists with updated disease and clinical information, classify them according to the genetic syndrome and the phenotypic location categories, and statistically identify the most relevant categories in each gene class. GFINDer is available for non-profit use at http://www.bioinformatics.polimi.it/GFINDer/.

INTRODUCTION

Remarkable improvements in bio-nano-technologies and biomolecular techniques have led to the increased production of experimental data that are rapidly accumulating in numerous and widely distributed heterogeneous databanks (1). Simultaneous development of information and communication technologies has enabled the efficient storing and easy retrieval of such data through the Internet. Now, in the post-genomic era, the challenge is developing methods to integrate the available data in order to comprehensively query them for extracting information leading to new biomedical knowledge (2,3). To this aim, the availability of biomolecular and biomedical ontologies and controlled vocabularies is of paramount importance to have common and standardized descriptions of concepts that enable to homogeneously classify data from heterogeneous sources (4,5). In addition, application of analysis and visualization techniques is essential to summarize data and highlight the most relevant information (6).

In the past few years several approaches have been developed for gene and gene product analyses, which provide valuable insights into gene relationships and protein interactions within specific biochemical pathways. Fewer computational contributions to phenotype analyses, aiming to unveil the complex molecular processes underlying phenotypic similar diseases, are yet to be provided. Besides clear intrinsic difficulties, one of the main reasons is the lack of access to controlled clinical information and its availability in structured form suitable for computational genome-wise analyses.

To enable performing comprehensive evaluations of functional gene annotations sparsely available in numerous different databanks accessible via the Internet, we previously developed GFINDer (7), a web server that dynamically aggregates functional annotations of user uploaded gene lists and allows performing their statistical analysis and mining. To this aim, GFINDer is organized in independent and interconnected modules that exploit several controlled vocabularies describing gene-related biomolecular processes and functions.

Here, we describe new original GFINDer modules specifically devoted to the analysis of genetic diseases and phenotypes. They exploit data from the OMIM databank (8,9) to allow annotating large numbers of user-classified biomolecular sequence identifiers with morbidity and clinical information, classifying them according to the related genetic diseases

*To whom correspondence should be addressed. Tel: +39 02 2399 3336/3303; Fax: +39 02 2399 3360; Email: masseroli@biomed.polimi.it and their phenotype locations (i.e. anatomical organ systems or types of findings), and statistically analyzing the obtained classifications. Such analyses can provide support for a phenotypic taxonomy of inherited diseases and facilitate a genomic approach to the understanding of fundamental biological processes and complex cellular mechanisms underlying patho-physiological phenotypes.

MATERIALS AND METHODS

Data

As a source of information on genetic diseases and their related phenotypes we used the OMIM databank, a comprehensive, authoritative and timely compendium of information in human genetics (8,9), currently containing 16 062 detailed entries about human genes and genetic disorders. Our main data source was the omim.txt file, which contains the entire free text of the OMIM databank. In addition to information on genetic loci, inheritance patterns and allelic variants, many OMIM entries contain a Clinical Synopsis section that delineates the accompanying signs and symptoms (i.e. phenotypes) of a disease and their locations. The Clinical Synopsis section is divided into phenotype location categories, either by organ system (e.g. cardiovascular, genitourinary and neurological) or by type of finding (e.g. inheritance and laboratory values).

To find the genes or genetic loci, if any that are involved in a disease we used the OMIM's morbidmap and considered the MIM codes associated with a gene, as provided by the Entrez Gene database (10).

Technologies and techniques

As previously performed for the first release of GFINDer web server (7), using information technologies that allow managing and analyzing a vast quantity of biological data with a simple user interface, we developed new GFINDer modules specifically devoted to genetic disorder analyses that enable performing exploratory and statistical morbidity and phenotypic evaluations of user-classified biomolecular sequence data. Using Java programming language, we implemented automatic procedures able to retrieve and keep updated in the GFINDer data tier genetic disorder information and correspondent gene annotations, as soon as new releases of them become available in the Entrez Gene and OMIM databanks. Genetic disease and phenotype data are automatically imported in a specifically designed MySQL relational database, where they are organized in hierarchical structures exploiting the multi-granular level and topological descriptions of OMIM morbidmap diseases and Clinical Synopsis phenotype locations, respectively.

Owing to the diversity of presentation of human diseases, and also possibly because OMIM has been in the development for decades, information in the OMIM Clinical Synopsis sections is not always represented in a uniform manner. Several typing errors and synonyms for the same name, and different names for overlapping concepts are often present for phenotype location categories (e.g. 'Endocrine' and 'Endocrine features', or 'Growth' and 'Development'), as well as for specific phenotypes (e.g. 'Aortic valve disease' and 'Aortic valvular disease', or 'Mental retardation, profound' and 'Mental retardation, severe'), which additionally include a high number of complex and specific findings (∼22 000). Therefore, we first extracted names of phenotypes and their location categories from Clinical Synopsis sections of omim.txt file by using text-parsing procedures. Then, we normalized the isolated phenotype location names by visually inspecting them and assigning a unique term to each synonym or incorrect written name. Finally, in GFINDer we implemented categorical analyses of genetic diseases and their phenotype locations based on the obtained list of unique category terms. In this list we did not consider the Clinical Synopsis 'Miscellaneous' and 'Molecular Basis' categories because they only contain free text descriptions and do not refer to truly observable phenotypes.

In the processing tier of our GFINDer web server, we implemented management and analysis procedures in Javascript and Active Server Page scripts and used Microsoft ActiveX Data Object technology and Standard Query Language to communicate with the MySQL DBMS server on the data tier. The statistical analysis routines we created employ hypergeometric and binomial distribution tests (11) and the Fisher's exact test (12) to assess statistical significance of the over and under representation of categorical biomedical and clinical annotations in a group of user-classified genes (7). Furthermore, different types of corrections for multiple tests (13) have also been included.

Using Hyper Text Markup Language and Javascript, a web graphic interface has been implemented for the GFINDer user layer, which is composed of any client computer connected to the web server on the processing layer through an Internet/Intranet communication network and loading in its client web browser the GFINDer graphic user interface.

RESULTS

Genetic disorder data normalization and structuring

Of the ∼16 000 entries in OMIM, we found 3805 that corresponded to genetic diseases and 4545 that contained a Clinical Synopsis section in which we initially found 135 different phenotype location category names. After name normalization, we had 93 unique category terms and 42 synonyms for 32 category names. Thus, for example, we combined 'Immune', 'Immunol', 'Immunologic' and 'Immunology' categories into a single 'Immunology' category, 'Metabolism' and 'Metabolic' into the only 'Metabolic' category, and merged each of the 'Ears', 'Eyes', 'Limbs', 'Lungs', 'Muscles' and 'Joints' categories with its correspondent singular term category, respectively. Total phenotype location entries were ∼32 554, whereas specific phenotype entries were 48 362.

Each description, of the nearly 3800 genetic diseases considered, was divided into up to four hierarchical levels according to its increasing degree of detail. Thus, for example, the OMIM disease descriptions 'Osteoporosis, idiopathic', 'Diabetes mellitus, insulin-resistant, with acanthosis nigricans' and 'Hypertension, early-onset, autosomal dominant, with exacerbation in pregnancy' were parsed and hierarchically structured in two, three and four hierarchical levels, respectively, as shown in Table 1. Similarly, the 93 unique phenotype locations were structured in three hierarchical levels according to their controlled topological descriptions (Table 2).

Table 1. Hierarchical structure of some of the genetic disease categories considered in the GFINDer, as derived from the correspondent disease descriptions provided by the OMIM databank

Genetic disease
Argininemia
Bipolar affective disorder
Diabetes mellitus
insulin-resistant
with acanthosis nigricans
Dysfibrinogenemia
alpha type
causing recurrent thrombosis
Epilepsy
partial
with auditory features
Glutaricaciduria
type I
Hypertension
early-onset
autosomal dominant
with exacerbation in pregnancy
Osteoporosis
idiopathic
Parkinson disease
juvenile
type 2
Tourette syndrome

Table 2. Defined hierarchical structure for some of the categories considered in the GFINDer to describe locations of genetic disease phenotypes, as partially provided by the OMIM databank and completed according to the anatomical organization of the described location categories

Phenotype location
Abdomen
Biliary tract
External features
Gastrointestinal
Liver
Pancreas
Spleen
Cardiovascular
Cardiac
Heart
Vascular
Genitourinary
Bladder
External genitalia
External genitalia, female
External genitalia, male
Internal genitalia
Internal genitalia, female
Internal genitalia, male
Kidneys
Ureters
Neurologic
Behavioral/psychiatric manifestations
Central nervous system
Peripheral nervous system

Such hierarchical structure, partially and inconsistently provided in the OMIM Clinical Synopsis sections, after normalization of phenotype location terms was completed and homogenized according to the anatomical organization of the described location categories. The main level of the three defined hierarchical levels, which includes broader organ systems or sites, comprised 36 locations. During gene list evaluations, the two defined disease and phenotype location hierarchical structures give each GFINDer user the chance to analyze specific diseases also aggregated into common general classes of syndromes, or particular phenotype locations also pooled into broader anatomical sites. This enables increasing the number of considered genes within more general disease and location categories, and thus allows unveiling their possible significance.

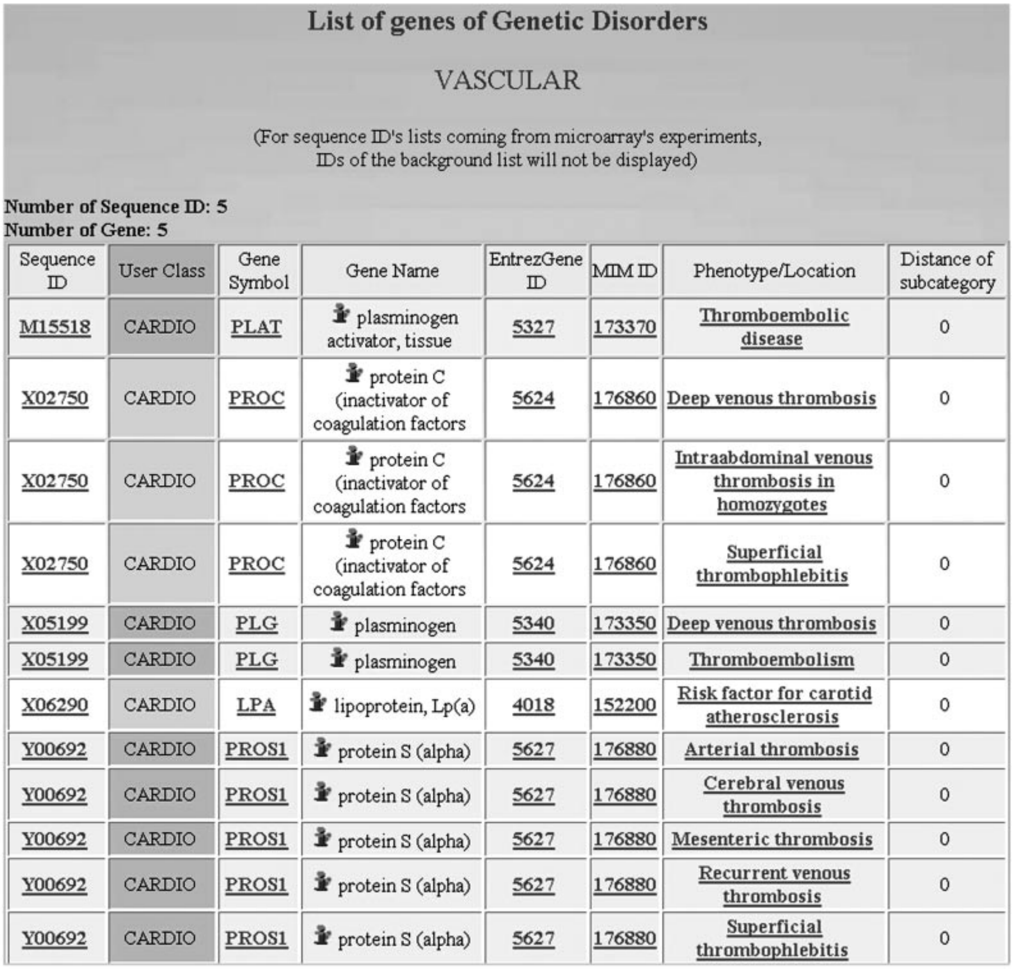
Web interface and genetic disorder modules

GFINDer user interface is organized in modules that allow to easily generate functional profiles of user-uploaded genes and evaluate their most significant characteristics through graphical views and statistical indexes in a web browser environment. The main modules for the analysis of genetic disorders related to user-selected genes are described in the following sections.

Upload and annotation. Through the Upload module, users can input a list of genes in the GFINDer web server (e.g. genes selected by means of microarray experiments and specified by GenBank accession numbers, RefSeq IDs, UniGene cluster IDs, Entrez Gene IDs, Affymetrix probe IDs or official gene symbols). In the list, each gene can appear grouped within predefined classes identified by any symbol (e.g. 1, −1 and 0; CLASS1 and CLASS2). For example, these classes can represent either gene expression regulations obtained from microarray experiments, or user classifications resulting from any clustering method, or different experimental biological conditions. The system automatically recognizes genes of several species, including *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Caenorhabditis elegans*, *Danio rerio* and *Drosophila melanogaster*.

The Annotation module enables to produce a tabular output of the uploaded gene list enriched with related genetic disease names, their inheritance mode and OMIM phenotype code, and with several other annotations, including official gene names and symbols, cytogenetic localizations, Entrez Gene identifiers, protein product identifiers and Gene Ontology categories with their evidence. These annotations, and all those provided within a window that opens by clicking on an annotated gene name (7), are automatically retrieved from the OMIM (9), Entrez Gene (10) and many other different databanks, including Gene Ontology (14), EBI-EMBL (15), KEGG (16), Swiss-Prot (17) and NetAffx (18).

Exploration. The Exploration Genetic Disorders module exploits the hierarchical structure defined for the considered OMIM data to perform evaluations on the genetic diseases and phenotype locations the genes in the user-uploaded list are involved in. By choosing the hierarchical level of detail (low levels provide lower specificity but higher coverage; high levels lead to lower coverage but to higher specificity), the module shows the genetic disease or phenotype location categories related to the input gene list, from the root down to the specified level of the data hierarchical structure. For each category, the category name, the absolute and percentage number of genes in the input list that are associated with the category, the list of these genes (Figure 1) and external links to the OMIM and other databank websites are provided. Within this module, a histogram graphical representation of



List of genes of Genetic Disorders

VASCULAR

(For sequence ID's lists coming from microarray's experiments, IDs of the background list will not be displayed)

Number of Sequence ID: 5
Number of Gene: 5

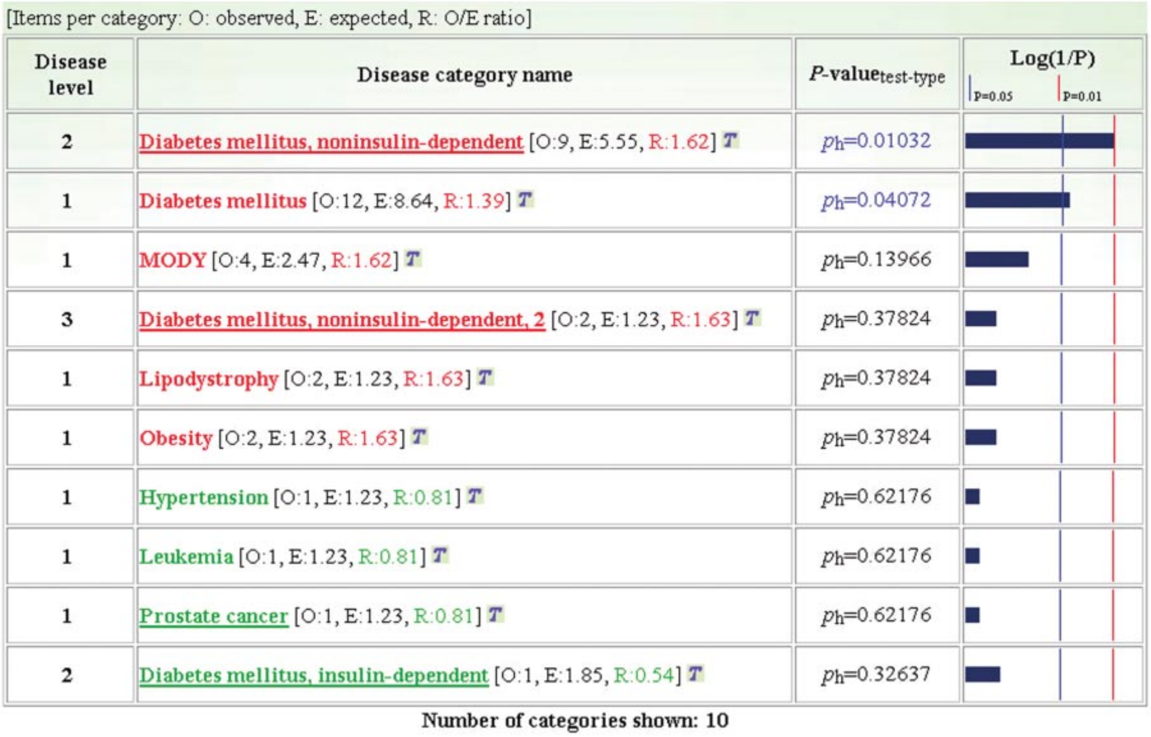
Sequence ID	User Class	Gene Symbol	Gene Name	EntrezGene ID	MIM ID	Phenotype/Location	Distance of subcategory
M15518	CARDIO	PLAT	plasminogen activator, tissue	5327	173370	Thromboembolic disease	0
X02750	CARDIO	PROC	protein C (inactivator of coagulation factors	5624	176860	Deep venous thrombosis	0
X02750	CARDIO	PROC	protein C (inactivator of coagulation factors	5624	176860	Intraabdominal venous thrombosis in homozygotes	0
X02750	CARDIO	PROC	protein C (inactivator of coagulation factors	5624	176860	Superficial thrombophlebitis	0
X05199	CARDIO	PLG	plasminogen	5340	173350	Deep venous thrombosis	0
X05199	CARDIO	PLG	plasminogen	5340	173350	Thromboembolism	0
X06290	CARDIO	LPA	lipoprotein, Lp(a)	4018	152200	Risk factor for carotid atherosclerosis	0
Y00692	CARDIO	PROS1	protein S (alpha)	5627	176880	Arterial thrombosis	0
Y00692	CARDIO	PROS1	protein S (alpha)	5627	176880	Cerebral venous thrombosis	0
Y00692	CARDIO	PROS1	protein S (alpha)	5627	176880	Mesenteric thrombosis	0
Y00692	CARDIO	PROS1	protein S (alpha)	5627	176880	Recurrent venous thrombosis	0
Y00692	CARDIO	PROS1	protein S (alpha)	5627	176880	Superficial thrombophlebitis	0

Figure 1. List of some of the input genes involved in genetic disorders with phenotype manifestations located in the vascular system. User class: user classification of uploaded sequence IDs; Phenotype/Location: phenotype, or its location, associated with a given sequence ID; Distance of subcategory: distance in the Clinical Synopsis hierarchy between a given phenotype, or location, level and the Vascular location level, which is associated with all genes of this list.

the distribution of the genes in the uploaded list among the related genetic diseases or phenotype locations is also given. Furthermore, for each phenotype locations a 'view' link enables to explore the particular phenotypic signs and symptoms occurring in that location, or in its sub-locations, for the uploaded genes. Histogram distribution, number, percentage and list of these genes involved in each specific phenotype are provided together with external links to the OMIM Clinical Synopsis section of such genes. Therefore, this module enables to easily and graphically understand either how many and which diseases, phenotype locations and their specific phenotypic signs and symptoms are correlated to each of the considered genes, or how many of the selected genes refer to each disease, location or phenotype, providing also several additional annotations on each gene.

Statistics. When in the uploaded input list genes are grouped in classes or a reference gene list is also loaded (e.g. the list of all the genes in the microarray used to produce the uploaded list of genes to analyze), now GFINDer allows performing statistical analyses on the genetic disorder categorizations of the input genes as well as on their Gene Ontology, biochemical pathway and protein domain functional categories. This enables highlighting in which of the considered functional and clinical categories the genes in the whole input list, or in each class contained, are involved and with what probability. Thus, a plain list of genes is enriched with biomedical meaning and statistical significances.

The Statistics Genetic Disorders module enables to estimate the relevance of the OMIM controlled annotations available for the input gene list, defining the level of detail to be considered for the genetic diseases or the phenotype locations in the data hierarchical structure created. To this aim, the annotated genes are grouped according to their class and disease or phenotype location categories, and their distribution among the considered categories is statistically evaluated by using hypergeometric or binomial distribution tests or the Fisher's exact test, as illustrated previously (7,11,12). After selecting a specific gene class, the module automatically considers each disease or phenotype location category represented in that class and provides a result table containing the observed number of input genes, their expected number, and the significance P-value of each category in the selected class with its histogram representation (Figure 2). Furthermore, for each phenotype location a link enables to statistically evaluate the particular phenotypic signs and symptoms occurring in



[Items per category: O: observed, E: expected, R: O/E ratio]

Disease level	Disease category name	*P*-value$_{\text{test-type}}$	Log(1/P)
			P=0.05 P=0.01

2	Diabetes mellitus, noninsulin-dependent [O:9, E:5.55, R:1.62] T	p_h=0.01032	
1	Diabetes mellitus [O:12, E:8.64, R:1.39] T	p_h=0.04072	
1	MODY [O:4, E:2.47, R:1.62] T	p_h=0.13966	
3	Diabetes mellitus, noninsulin-dependent, 2 [O:2, E:1.23, R:1.63] T	p_h=0.37824	
1	Lipodystrophy [O:2, E:1.23, R:1.63] T	p_h=0.37824	
1	Obesity [O:2, E:1.23, R:1.63] T	p_h=0.37824	
1	Hypertension [O:1, E:1.23, R:0.81] T	p_h=0.62176	
1	Leukemia [O:1, E:1.23, R:0.81] T	p_h=0.62176	
1	Prostate cancer [O:1, E:1.23, R:0.81] T	p_h=0.62176	
2	Diabetes mellitus, insulin-dependent [O:1, E:1.85, R:0.54] T	p_h=0.32637	

Number of categories shown: 10

Figure 2. GFINDer statistical analysis of inherited diseases associated with the considered, non-insulin dependent diabetes mellitus versus insulin dependent diabetes mellitus related genes. Disease level: level in the defined Disease hierarchy of a given disease (higher levels correspond to more detailed and specific diseases); *P*-value$_{\text{test-type}}$: *P*-value defining association between a given disease and a considered class of genes, and the initial of used statistical test name (h: hypergeometric distribution test).

that location, or in its sub-locations, for the uploaded genes. Observed and expected numbers of these genes involved in each specific phenotype are provided together with significance *P*-value histogram distribution of particular signs and symptoms in that location, or its sub-locations, for the uploaded genes in the selected class. External links to the OMIM Clinical Synopsis sections of these genes are also given.

Validation and applications

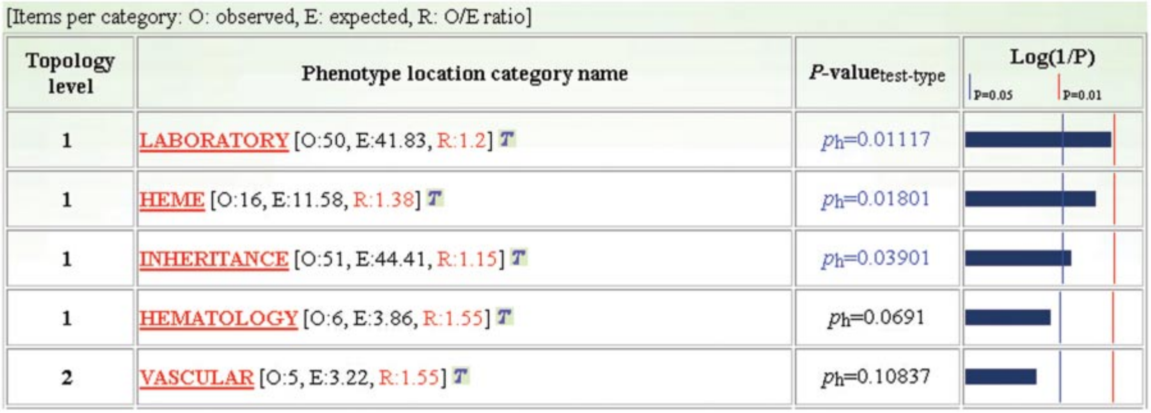
To demonstrate GFINDer's potential, we used it to functionally analyze two sets of genes. The first set was composed of 107 Entrez Gene IDs of diabetes-related genes involved either in insulin dependent diabetes mellitus (IDDM) (41 genes) or in non-insulin dependent diabetes mellitus (NIDDM) (66 genes), as described in some review articles on the subject (19,20). The second set of 1046 genes was identified by GenBank accession numbers of distinct human clones spotted on the 7734-1 or 7736-1 Clonetech microarrays (BD Biosciences ClonTech; http://www.bdbiosciences.com/clontech/; last access: April 15, 2005), focused on cardiovascular system (522 clones) and neurobiology (524 clones), respectively.

GFINDer annotation of the first set of genes showed that for 43 of them (41 autosomal and 2 X-linked) relations to 89 inherited diseases were known. Thus, we used the statistical Genetic Disorders module to evaluate the relevant presence, in the NIDDM versus IDDM class, of genes involved in specific genetic diseases. We concentrated on those disease categories associated with at least two of the selected 107 genes. The most relevant categories that the statistical analysis highlighted were in agreement with the diseases in which each group of the selected genes is known to be involved (Figure 2). In fact, they included 'Diabetes mellitus, noninsulin-dependent' (*P* = 0.01032), 'Diabetes mellitus' (*P* = 0.04072) and 'MODY' (*P* = 0.13966) for the NIDDM gene class, and 'Diabetes mellitus, insulin-dependent' (*P* = 0.32637) for the IDDM class. Furthermore, disease categories with higher *P*-values indicated disorders that can be associated with diabetes.

Using the statistical Genetic Disorders module, we also analyzed the second considered set of genes. We statistically evaluated the phenotype locations of known inherited diseases most represented in the considered cardiovascular system (CARDIO) related genes versus the neurobiology (NEURO) associated genes. As illustrated in Figure 3, GFINDer analysis results correctly highlighted phenotype locations of signs and symptoms related to the cardiovascular system rather than to neurobiology. Among others, such phenotypes included 'Bleeding diathesis' and 'Hemorrhagic diathesis' in the 'Heme' location category, 'Deep venous thrombosis' and 'Superficial thrombophlebitis' in the 'Vascular' category, and 'Anemia' in the 'Hematology' category. All these findings validate our approach developed for the analysis of genetic disorder related genes, implemented and made available through the GFINDer web server.

DISCUSSION

In the past few years, several tools [e.g. DAVID (21), FatiGO (22), GFINDer (7), GoMiner (23), GOTM (24), NetAffx (18)



[Items per category: O: observed, E: expected, R: O/E ratio]

Topology level	Phenotype location category name	P-value$_{\text{test-type}}$	Log(1/P)
1	LABORATORY [O:50, E:41.83, R:1.2] T	p_h=0.01117	
1	HEME [O:16, E:11.58, R:1.38] T	p_h=0.01801	
1	INHERITANCE [O:51, E:44.41, R:1.15] T	p_h=0.03901	
1	HEMATOLOGY [O:6, E:3.86, R:1.55] T	p_h=0.0691	
2	VASCULAR [O:5, E:3.22, R:1.55] T	p_h=0.10837	

Figure 3. GFINDer statistical analysis of genetic disorders: the phenotype location categories most significantly over-represented in the considered cardiovascular system versus neurobiology related genes. Topology level: level in the defined Clinical Synopsis hierarchy of a given phenotype location (higher levels correspond to more detailed and specific locations); P-value$_{\text{test-type}}$: P-value defining association between a given phenotype location and a considered class of genes, and the initial of used statistical test name (h: hypergeometric distribution test).

and Onto-Tools (25)] have been proposed to enrich lists of genes with biological information and estimate the type of information more relevant for a given set of genes. Most of these tools only use the Gene Ontology controlled vocabularies (14) as sources of categorical information to analyse. Very few of them (i.e. DAVID and GFINDer) also enable analyzing other controlled vocabularies (e.g. the KEGG biochemical pathways) but to our knowledge, at present, none allows phenotype analyses.

Understanding clinical phenotypes through their corresponding genotypes may be relatively simple with single gene syndromes. Nevertheless, more complex diseases often consisting of different clinical phenotypes that may result from interactions among multiple and potentially unknown genetic loci, require more complicated, multivariate methods of analysis (26). Furthermore, often the same phenotype may be caused by significantly different genetic alterations. Thus, analyzing the molecular processes underlying phenotypically similar diseases may provide insight into these complex interactions (27). Tools for analyzing the large amounts of biological data being produced are the key to understand these intricate relationships. We agree that the analysis of phenotypic similarities among diverse diseases associated with known loci may provide insight into the genetic interactions underlying them and could ultimately give clinically useful insights into disease processes, including more complex diseases influenced by multiple genetic loci (26).

OMIM is a very valuable, ample and freely accessible resource of information on known genetic diseases and their phenotypes. Unfortunately, it is mainly constituted of textual descriptions organized for interactive browsing rather than automated analyses. Our efforts enabled to structure some of the OMIM information in a way suitable for computational use and led to the implementation of the new GFINDer modules for the analysis of genetic disorders related genes. As our validation results showed, these modules correctly highlight the genetic diseases and their phenotypic locations significantly more represented within user-defined classes of genes, independently on the methods used to define them. Therefore, they represent a uniquely valuable tool for the numerous users of the GFINDer web server (which has had ∼26 000 accesses from more than 800 distinct IPs from many countries since its release in 2004). In fact, supporting both biomolecular functional evaluations and phenotype analyses of inherited diseases, GFINDer facilitates a genomic approach to the understanding of fundamental biological processes and complex cellular mechanisms underlying patho-physiological phenotypes. However, it is important to note that the annotations and analyses provided by the GFINDer can only be as accurate as the underlining online sources from which the annotations are retrieved. The GFINDer web server is freely available online for non-profit use at http://www.bioinformatics.polimi.it/GFINDer/.

ACKNOWLEDGEMENTS

We thank Bisharah Libbus for selecting from review articles the list of genes known to be most relevantly involved in diabetes that we used to validate GFINDer Genetic Disorders modules, and Karina Gibert for her support in tuning and improving GFINDer statistical modules. The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Galperin,M.Y. (2005) The Molecular Biology Database Collection: 2005 update. *Nucleic Acids Res*., **33**, D5–D24.
2. Kanehisa,M. and Bork,P. (2003) Bioinformatics in the post-sequence era. *Nature Genet*., **33**, S305–S310.
3. Rhodes,D.R. and Chinnaiyan,A.M. (2004) Bioinformatics strategies for translating genome-wide expression analyses into clinically useful cancer markers. *Ann. NY Acad. Sci*., **1020**, 32–40.
4. Schulze-Kremer,S. (2002) Ontologies for molecular biology and bioinformatics. *In Silico Biol*., **2**, 179–193.
5. Stevens,R., Goble,C.A. and Bechhofer,S. (2000) Ontology-based knowledge representation for bioinformatics. *Brief Bioinformatics*, **1**, 398–414.
6. Soukup,T. and Davidson,I. (2002) *Visual Data Mining: Techniques and Tools for Data Visualization and Mining*. John Wiley & Sons, NY.

7. Masseroli,M., Martucci,D. and Pinciroli,F. (2004) GFINDer: genome function integrated discoverer through dynamic annotation, statistical analysis, and mining. *Nucleic Acids Res.*, **32**, W293–W300.
8. McKusick,V.A. (1998) *Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders, 12th edn*. John Hopkins University Press, Baltimore, MD.
9. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.
10. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **33**, D54–D58.
11. Casella,G. and Berger,R.L. (2002) *Statistical Inference, 2nd edn*. Duxbury Press, Belmont, CA.
12. Fisher,L.D. and van Belle,G. (1993) *Biostatistics: A Methodology for the Health Sciences*. John Wiley & Sons, NY.
13. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, **57**, 289–300.
14. The Gene Ontology Consortium (2000), Gene Ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
15. Kanz,C., Aldebert,P., Althorpe,N., Baker,W., Baldwin,A., Bates,K., Browne,P., van den Broek,A., Castro,M., Cochrane,G. *et al.* (2005) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **33**, D29–D33.
16. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
17. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
18. Liu,G., Loraine,A.E., Shigeta,R., Cline,M., Cheng,J., Valmeekam,V., Sun,S., Kulp,D. and Siani-Rose,M.A. (2003) NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res.*, **31**, 82–86.
19. Adeghate,E. (2004) Molecular and cellular basis of the aetiology and management of diabetic cardiomyopathy: a short review. *Mol. Cell. Biochem.*, **261**, 187–191.
20. Xu,R., Li,H., Tse,L.Y., Kung,H.F., Lu,H. and Lam,KS. (2003) Diabetes gene therapy: potential and challenges. *Curr. Gene Ther.*, **3**, 65–82.
21. Dennis,G.,Jr, Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C. and Lempicki,R.A. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.*, **4**, P3.
22. Al-Shahrour,A.F., Díaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.
23. Zeeberg,B.R., Feng,W., Wang,G., Wang,M.D., Fojo,A.T., Sunshine,M., Narasimhan,S., Kane,D.W., Reinhold,W.C., Lababidi,S. *et al.* (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**, R28.
24. Zhang,B., Schmoyer,D., Kirov,S. and Snoddy,J. (2004) GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using gene ontology hierarchies. *BMC Bioinformatics*, **5**, 16.
25. Khatri,P., Bhavsar,P., Bawa,G. and Draghici,S. (2004) Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments. *Nucleic Acids Res.*, **32**, W449–W456.
26. Phillips,T.J. and Belknap,J.K. (2002) Complex-trait genetics: emergence of multivariate strategies. *Nature Rev. Neurosci.*, **3**, 478–485.
27. Cantor,M.N. and Lussier,Y.A. (2004) Mining OMIM for insight into complex diseases. *Medinfo*, **2004**, 753–757.